Intelética 2(4) (2025), 135-149
Revista de Inteligencia Artificial, Ética y Sociedad

INTELETICA

https://inteletica.iberamia.org/

# AI, Human, or Hybrid?
# Reliability of AI Detection Tools in Multi-Authored Texts

Sheila Queralt[1] , Beatriz Esparcia[2] , Marco R. Lessi[3], Lucía Sánchez-Vecina[4] y Laura C. Úbeda-Cuspinera[5]

[1] Laboratorio SQ-Lingüistas Forenses
[2] Laboratorio SQ-Lingüistas Forenses
[3] Universitat Autònoma de Barcelona
[4] Laboratorio SQ-Lingüistas Forenses
[5] Universitat de Barcelona

[1] sheila.queralt@cllicenciats.cat
[2] esparcia.beatriz@gmail.com
[3] marcoriccardo.lessi@uab.cat
[4] lucia.sv.trad@gmail.com
[5] lubeda@ub.edu

**Abstract.** This article presents the first results of the CorpIdentIA project (Corpus Identity & Authorship Intelligence Analysis), focused on the analysis of texts generated wholly or partially by artificial intelligence. Based on an experimental Spanish corpus (n = 180) that includes human, artificial, and mixed texts, the study analyzes the performance of three detectors (Originality.ai, GPTZero, and Copyleaks) against different generative models (ChatGPT, Gemini, and Grok). The main objective of this study is to evaluate the effectiveness of different AI detection tools in classifying these texts according to their origin (AI, human, or hybrid). The results reveal significant differences among tools: Originality.ai shows the best overall performance, while GPTZero stands out for its low rate of false positives. However, none of the tools demonstrates acceptable reliability in detecting hybrid texts. Recurrent biases are observed depending on the AI model, along with misclassifications with high confidence, which raises risks in the implementation of these tools without expert human review. This work contributes to the current debate on the trustworthiness of detectors, the risk of false accusations in forensic contexts, and the need for explainable approaches from applied linguistics. Besides, these findings underline the importance of interdisciplinary collaboration between linguists, computer scientists, and legal experts.

**Resumen.** Este artículo presenta los primeros resultados del proyecto CorpIdentIA (Corpus Identity & Authorship Intelligence Analysis), centrado en el análisis de textos generados total o parcialmente por inteligencia artificial. A partir de un corpus experimental en español (n = 180) que incluye textos humanos, artificiales y mixtos, el estudio analiza el rendimiento de tres detectores (Originality.ai, GPTZero y Copyleaks) frente a distintos modelos generativos (ChatGPT, Gemini y Grok). El objetivo principal de este estudio es evaluar la eficacia de diferentes herramientas de detección de IA en la clasificación de estos textos según su origen (IA, humano o híbrido). Los resultados revelan diferencias significativas entre las herramientas: Originality.ai muestra el mejor rendimiento global, mientras que GPTZero destaca por su bajo índice de falsos positivos. Sin embargo, ninguna de las herramientas demuestra una fiabilidad aceptable en la detección de textos híbridos. Se observan sesgos recurrentes en función del modelo de IA, así como clasificaciones erróneas con alta confianza, lo que plantea riesgos en la implementación de estas herramientas sin revisión experta humana. Este trabajo contribuye al debate actual sobre

ISSN: 3020-7444

la fiabilidad de los detectores, el riesgo de falsas acusaciones en contextos forenses y la necesidad de enfoques explicables desde la lingüística aplicada. Asimismo, estos hallazgos subrayan la importancia de la colaboración interdisciplinar entre lingüistas, informáticos y juristas.

**Palabras clave**: estilística computacional, lingüística forense, atribución de autoría, discurso generado por máquina.

**Keywords**: computational stylistics, forensic linguistics, authorship attribution, machine-generated discourse

# 1 Introduction

The emergence of large language models (LLMs) has radically transformed textual production across multiple domains, from academic writing to institutional and professional communication. This expansion has given rise to increasing concern about the detection of texts generated—wholly or partially—by artificial intelligence (AI), especially in contexts where human authorship is essential: educational settings, judicial processes, evaluation of scientific publications, or the attribution of contractual responsibilities.

In response to this new reality, automatic AI detection tools have proliferated, claiming to distinguish between human and artificial texts with high levels of accuracy. However, numerous studies have warned of the technical and conceptual limitations of these detectors: linguistic biases, misclassification errors, difficulties with hybrid texts, and lack of transparency in their decision-making mechanisms [1, 2, 3, 4].

In particular, the identification of mixed texts—those combining human and artificial intelligence input—has emerged as one of the main challenges for current systems. Despite their increasing frequency in real practice, such texts are often forcibly categorized within binary labels (human or AI), overlooking the discursive complexity introduced by intermodal collaboration. This simplification can lead to significant attribution errors, which are especially serious in forensic or evaluative contexts where precision, explainability, and empirical evidence are required.

The main objective of this study is to assess the reliability of different AI text detection tools in realistic, multi-authored contexts, with particular attention to their performance on hybrid texts. Unlike previous studies focused exclusively on binary classification (human vs. AI), this work also examines intermediate cases, where authorship is shared or where there has been cross-intervention between human and artificial intelligence.

To this end, an experimental Spanish-language corpus has been designed, comprising controlled texts organized into four modalities: human (H), artificial (AI), human edited by AI (H + AI), and AI modified by humans (AI + H). The analysis focuses on three tools currently available to general users (Copyleaks, GPTZero, and Originality.ai), applied to this multivariate corpus.

Specifically, the study seeks to determine the accuracy and error rates of each tool across the different textual modalities; to evaluate their capacity not only to identify automatic authorship, but also the hybrid nature of texts; to analyze the impact of the generative model (ChatGPT, Gemini, and Grok) and type of intervention (partial or total) on the results; and, finally, to assess the real usefulness of these tools in forensic, educational, or institutional contexts, where authorship attribution carries significant consequences.

# 2 Theoretical Framework

## 2.1 Detection of AI-generated text: challenges and advances

Authorship attribution of a text can be defined as the process by which the linguistic features of a text are examined in order to draw conclusions about its authorship [5]. Authorship attribution is a field widely addressed in the academic literature. Some authors who have dealt with this issue include [6, 7, 8, 9, 10], among many others. The research lines of this discipline are expanding with the emergence of AI-generated text, since after the rise and proliferation of large language models, alarm bells have quickly been raised regarding the impact of these tools. On the one hand, the potential of these language models is evident, as well as the major progress they may bring in fields such as medicine, education, law, programming, and journalism, among many others.

Nevertheless, it is also worth mentioning the inherent danger posed by the fraudulent use of these tools, especially in fields such as education [11], where students may claim synthetically generated texts as their own. Both major academic institutions [12] and educators [13, 14, 15] have expressed concern regarding these new tools. Likewise, the ethical issues raised by the use of such applications have also been highlighted [16, 17].

In light of these concerns, determining whether a text has been generated by AI has become one of the main objectives of natural language processing. At an initial stage, it was suggested that LLMs themselves could identify whether a text is synthetic or human [18, 19], or that restricted-use tools and workshops could be developed which remained unnoticed beyond the scientific community, such as the AuTexTification [20] or SemEval-2024 tasks [21].

Currently, as a response to the fears arising from the widespread use of LLMs, a wide range of freely accessible or commercial tools proliferate online, claiming to be capable of identifying whether a text has been artificially generated. Comparisons between LLM models and whether their content can be detected are subjects of broad debate in the specialized literature on the matter. Some authors who have addressed this issue include [22, 23, 24, 25, 26, 27, 28], among others. However, these tools have been subjected to studies assessing their real effectiveness, yielding mixed results. In general, the main difficulty faced by such automatic detectors is not distinguishing between human and synthetic texts, but rather a third scenario: hybrid texts—that is, those generated by AI and modified by humans, or vice versa [29].

## 2.2 Previous studies on the evaluation of automatic detectors: biases and limitations

In recent years, the evaluation of automatic AI detectors has attracted the interest and analysis of the scientific community. Some tools and models have attempted to determine whether a text is human by focusing on the features of this type of writing [30, 31, 32]. However, in general terms, these applications reveal the opposite tendency: to evaluate stylometric, lexical, and syntactic features characteristic of AI-generated language.

For their part, the companies responsible for these LLMs also seem unable to clarify their functioning. The developers of ChatGPT themselves designed an open-access tool for detecting content generated by their application [33], although in the end they were unable to reliably determine whether the content was synthetic or not [34].

At present, numerous studies and tests have been conducted with AI detection tools to evaluate their performance. Many works compare different detectors with each other or against proprietary tools, as in the case of studies presented by [1, 2, 3, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44], among many others.

Nevertheless, although these studies show diverse quantitative results, two limitations are systematically highlighted in the conclusions of works addressing the evaluation of AI detectors. First, a bias towards certain languages appears to be observed, with detectors generally showing very poor performance in multilingual contexts [4]. In addition, detectors tend to perform well in binary classification tasks—that is, they are effective in labeling texts as either entirely human or entirely synthetic—but their performance decreases considerably when they must identify so-called hybrid texts [29, 45, 46].

## 2.3 Typologies of text generation: human, artificial and mixed

This section examines the main features that differentiate human, artificial, and mixed texts. One of the most relevant issues identified for text classification is that automatic identification tasks are usually based on dichotomous criteria (human vs. artificial), without considering a possible third scenario (mixed or hybrid) [47].

Likewise, few works were found in the reviewed literature in which the participants (if any) tasked with discerning whether a text was human or AI-generated were linguists[1] [16]. According to this study, the participants identified a series of characteristics that, in their view, could guide the decision on the authorship of a text. These include: the continuity and coherence of the text, specificity or vagueness in detail, familiarity and voice, and the quality of writing at the sentence level. More specifically, some participants pointed out that human texts were easier to read and more understandable due to the logical connections established between ideas [16].

[1] In general, the issue of automatic detection has been addressed from a computational perspective.

Along similar lines, [48] point out that humans tend to produce texts that are less cognitively demanding, with greater semantic content and richer emotional expression, compared to synthetic texts.

For their part, [47] also establish a set of distinguishing features of human texts as opposed to automatic ones. Among these are linguistic, punctuation, paralinguistic, and psycholinguistic features.

Finally, it is worth noting that none of the reviewed studies provides a characterization of hybrid texts. Overall, this phenomenon continues to be addressed in binary terms rather than conceptualized along a continuum.

# 3 Methodology

This study is undertaken within the framework of the CorpIdentIA project (Corpus Identity & Authorship Intelligence Analysis), which has resulted in the development of an experimental methodology. Said methodology combines the design of an ad hoc corpus, the simulation of textual production scenarios, and the comparative evaluation of automatic tools for authorship and text origin detection. The methodology employed builds on previous work by [1, 39].

## 3.1 Corpus design and analyzed modalities

For the purposes of this study, an ad hoc corpus was compiled, named the Corpus CorpIdentIA, consisting of a total of 180 texts distributed across four distinct modalities of authorship and textual production. These modalities make it possible to simulate different real-world scenarios of content generation, following the classification into modalities proposed in previous works by [42, 44, 45, 49].

a) The Human Modality (H) comprises 30 texts written entirely by humans without AI involvement, selected from a work published before the widespread availability of generative models [50].

b) The AI Modality (AI) also includes 30 texts, generated exclusively by three state-of-the-art large language models (LLMs): 10 with ChatGPT, 10 with Gemini, and 10 with Grok.

c) The Human with AI Support Modality (H + AI) consists of another 30 texts, where the original human productions from section H were subsequently modified through AI-based editing or reformulation tools. Again, 10 texts were produced with ChatGPT, 10 with Gemini, and 10 with Grok.

d) The AI with Human Revision Modality (AI + H) is the largest subset, with 90 texts initially generated by the three LLMs and later revised by humans through three types of intervention: insertion of grammatical errors, addition of human-written sentences, and punctuation modifications. For each type of intervention, 10 texts per model (ChatGPT, Gemini, and Grok) were created, resulting in 30 texts per intervention.

In total, the corpus comprises 180 texts, evenly distributed across human, artificial, and hybrid modalities, with 10 texts per submodality in all cases.

## 3.2 Types of text and simulated scenarios

The design of the corpus follows the methodological principles established in [49], who proposed an empirical approach to the construction of forensic corpora through the simulation of realistic communicative situations in legal or police contexts. In that study, participants were asked to produce texts under six different scenarios, four of them related to medium–high level threats (Scenarios 1 to 4) and two concerning ordinary correspondence (Scenarios 5–6).

In the present work, consistent with that model, only Scenario 1 – *Bail and Walls* was selected, as it was considered representative of a plausible interpersonal conflict with potential legal implications. This decision responds to the need to homogenize the corpus and control discursive variables in order to accurately assess the differences between human, artificial, and mixed-origin texts.

The scenario was presented as follows to the participants in 2015 and, in the present project, to the AI tools:

*Some time ago, you rented a single room, for which you had to pay a deposit of 3,000 euros. You are now moving to a new apartment, and the former landlord refuses to return any part of the deposit, claiming that you left the room walls in poor condition and that they therefore need to be repainted. You have explained in a hundred different ways that this is not true and even shown before-and-after photos to compare the condition, but he insists that the walls are damaged and refuses to return the deposit.*

*You suggest different options, such as repainting the walls yourself, but no solution is reached. In the end, you give up a significant part of the deposit corresponding to the cost of paint and labor, even though the apartment is in perfect condition, because you believe this way you might recover at least part of the deposit. There is still no result, as the landlord continues to refuse to return a single cent. Finally, your only option is to write him a letter in which you threaten him in order to obtain your deposit.*

Based on this situation, participants were asked to produce a text of approximately 600 words, written in the first person, with a firm and emotional tone while avoiding explicit threats. The simulated author's sociolinguistic profile was defined as a young person (between 20 and 35 years old), born in Catalonia, with a university degree in Translation and Interpreting. In this project, this characterization was also included in the prompts used for artificial generation, with the aim of ensuring stylistic comparability across modalities and minimizing tonal or rhetorical bias:

*Write a first-person letter of approximately 600 words. The person writing is between 20 and 35 years old, was born in Catalonia, and has a university degree in Translation and Interpreting. The tone should be firm and emotional, aiming to convince or put pressure on the other party without making explicit threats. Use natural language appropriate for a young person with good linguistic competence.*

The choice of an average text length of 600 words is based on recommendations found in previous studies. [51] notes that qualitative studies can work with texts of 150–200 words, while quantitative approaches—particularly those employing stylometric techniques—require longer texts. [52], for example, showed that accuracy in authorship detection decreases significantly when texts are shorter than 1,000 words. In our case, an intermediate length was adopted, allowing for analysis with computational tools without compromising the linguistic richness of the data.

For the H texts modified by the three AI models, the same prompt was used: *Improve the text and explain what you have improved.*

## 3.3 Artificial generation: LLM tools used

The applications selected to carry out the project were the following:

### 3.3.1 Generation

For text generation, the tools employed were: ChatGPT (GPT-4o Mini), Gemini (version 2.5 Pro), and Grok (version 3). It should be noted that both ChatGPT and Gemini displayed moral considerations when asked to write or rewrite a threatening letter. With respect to Gemini, it was necessary to adopt several persuasion strategies to achieve this purpose. Among these strategies, it had to be stated that the objective of generating such texts was academic and research-oriented. ChatGPT, however, only objected in the creation of the H + AI texts, and it was sufficient to resend the prompt, to which the tool responded that it would modify the threatening tone of the texts. Grok, for its part, proved to be the most accessible model in terms of moral considerations.

The tools initially selected were those best ranked in the LLMArena[2] leaderboard under the text category. However, the preselected models (Gemini, ChatGPT, DeepSeek, Claude, and Grok) were subsequently evaluated to ensure adequate performance in Spanish in line with the objectives of this analysis. Based on this review, it was decided to exclude DeepSeek and Claude.

[2] https://lmarena.ai/leaderboard

#### 3.3.2 Detection

The tools selected for detection were the following:

First, Copyleaks was selected, a commercial text analysis tool that relies on deep neural networks to identify patterns present in AI-generated texts and the relationships among them. The results given by this tool indicate(s) the percentage of the text it considers to have been generated by artificial intelligence. It supports more than 30 languages (including Spanish) and detects popular LLMs such as ChatGPT, Gemini, or Claude. According to independent studies[3], its accuracy is above 99%.

Second, GPTZero is a detection tool based on deep neural networks, combined with a statistical model using the metrics of perplexity and burstiness. GPTZero provides a probability percentage that represents the level of confidence regarding AI involvement. The detector classifies texts as AI, H, or Mixed. However, the verbal labels for possible mixed classifications differ across the current models ("written by human and AI" and "created by human and edited by AI"), and the tool does not allow users to select the desired model, as it automatically assigns one in each detection based on unknown criteria. Moreover, the inversely directional label "created by AI and edited by human" is currently unavailable, which may indicate a potential bias in the opposite direction.

Third, Originality.ai [23, 24, 42] is a tool that uses modified versions of Google's deep neural network model BERT. According to internal studies published in early 2024, the model was trained on the most recent LLMs available at that time and achieved 98.8% true positives (TP). Subsequently, a new Multilanguage version was implemented, which was used for the present study, capable of detecting artificial text in up to 30 different languages, with metrics surpassing 98% accuracy in Spanish. This tool only provides global binary results based on sentence-level confidence: Likely AI and Likely Original, classifying texts respectively as artificial or human. Therefore, it does not consider any text as belonging to a mixed category.

### 3.4 Evaluation metrics: accuracy, error, bias, and confusion

The performance of the detection tools was evaluated through a mixed approach that combines standard binary classification metrics (accuracy, precision, recall, F1-score, false positive rate, and false negative rate) with an additional coding system tailored to the complexity of hybrid texts. In these cases, we distinguished true and false classifications not only for human and AI texts, but also for hybrid categories (TP-MIXED, FP-MIXED, FN-MIXED).

A specific methodological decision was made to ensure consistency across tools: whenever a detector identified hybrid intervention, the case was considered a true positive mixed (TP-MIXED), regardless of whether the system described the sequence as *H + AI* or *AI + H*. This decision reflects the main objective of the study: to evaluate whether the tools can reliably detect AI involvement in textual production, beyond the exact direction of the process.

This framework combines quantitative rigor with qualitative interpretation of system outputs, enabling a more accurate assessment of the actual performance of AI detectors in the forensic analysis of texts.

## 4 Results

The present study evaluated the performance of three AI text detection tools (Originality.ai, GPTZero, and Copyleaks) using a corpus of 180 Spanish texts, including human, artificial, and hybrid texts. Unlike many previous works focused on English as the main language [28, 30], this research stands among the first systematic studies in Spanish, contributing a more representative multilingual perspective.

### 4.1 Detection of entirely human and AI texts

This section analyzes the performance of automatic detection tools in the classification of texts written exclusively by humans or generated by artificial intelligence—that is, in a dichotomous scenario without hybrid intervention.

[3] For further information, see third-party studies on Copyleaks: https://copyleaks.com/blog/ai-detector-continues-top-accuracy-third-party

The corpus used for this test consisted of 60 texts: 30 produced by humans and 30 by LLM models (ChatGPT, Gemini, and Grok). The tools evaluated were GPTZero, Originality.ai, and Copyleaks.

The results show a clear difference in performance among the tools. Originality.ai achieved the best overall performance, with 100% accuracy on AI texts and 90% on human texts. GPTZero followed, with an accuracy rate of 96.7% for human texts and 86.7% for AI texts. Copyleaks ranked last, with much lower performance and an accuracy rate of 76.7% for both classes.

Figure 1 provides a visual representation of the confusion matrices for each tool, allowing direct observation of the proportions of TP, TN, FP, and FN. These values, fundamental for quantitative evaluation, form the basis for the calculation of more advanced metrics.

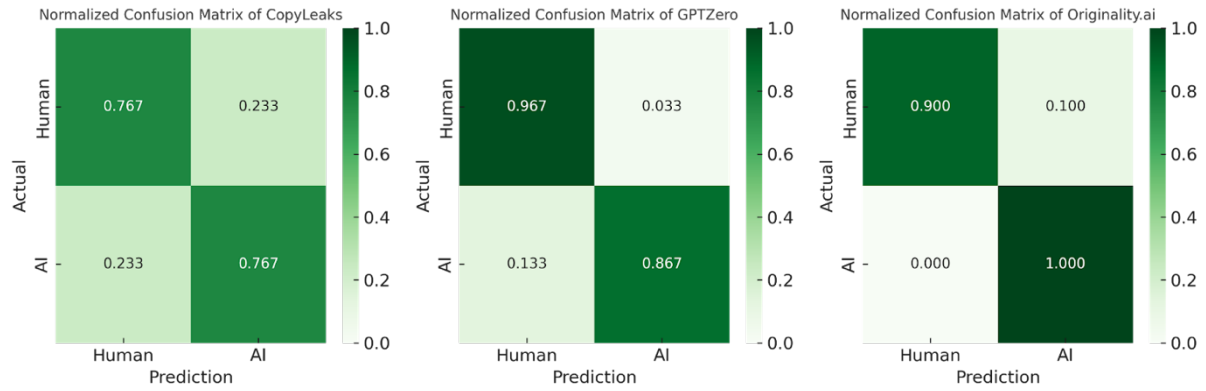


Figure 1. Visual comparison of normalized confusion matrices by detection tool.

Based on the results visually presented in the matrices in Figure 1, the following metrics were calculated to more accurately assess the performance of the tools: accuracy, precision, recall, F1-score, FPR, and FNR, in order to evaluate the performance of each tool with greater precision.

In the case of human texts, where the only possible error is misclassification as AI (FP), the FPR rates show clear differences between tools. GPTZero demonstrated the best performance, with only 3.3% errors (FPR = 0.0333), followed by Originality.ai with 10% (FPR = 0.1). Copyleaks ranked last, wrongly assigning AI involvement in 23.3% of human texts (FPR = 0.2333), representing the highest number of false positives among the three tools.

In the category of texts generated entirely by AI, the results show a different trend. Originality.ai achieved the best performance with an FN rate of 0.0, meaning it correctly identified all synthetic texts (recall = 1.0). GPTZero ranked second, with 4 such errors (FNR = 0.1333). Copyleaks, once again, showed the worst performance, with 7 AI texts not correctly detected, resulting in an FNR of 0.2333 (see Figure 1 and Table 2).

Figure 2 presents a visual comparison of three main metrics—precision, recall, and F1-score—in a radar chart. These metrics were selected because they are the most suitable for evaluating the overall performance of binary classification systems, as they allow assessment of both the proportion of correct classifications and the ability to correctly detect positive cases, which are key aspects in automatic detection tasks.

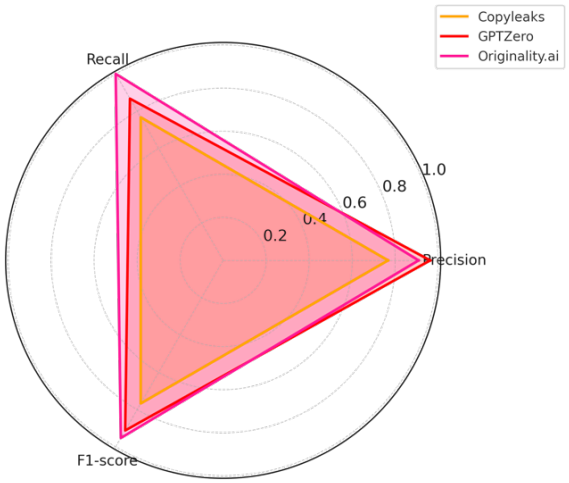


Figure 2. Radar chart comparing performance metrics on purely human or AI-generated texts.

The data confirm the strong performance of Originality.ai and the higher precision of GPTZero, although its recall decreases compared to Originality.ai. As for Copyleaks, a notable reduction in effectiveness is observed in comparison with the other two tools. Copyleaks is less precise and also has greater difficulty detecting AI-generated texts.

Beyond the intrinsic quality of each tool, one factor that appears to strongly condition performance is the AI generative model used. As shown in Figure 3, the recall of the detectors varies depending on the LLM. In particular, texts generated by Gemini turn out to be the most problematic for GPTZero and especially for Copyleaks, which makes 6 of its 7 FN errors with this model, reducing its recall to 40%. GPTZero also lowers its performance in this subcategory (recall: 80%), while Originality.ai maintains perfect coverage.

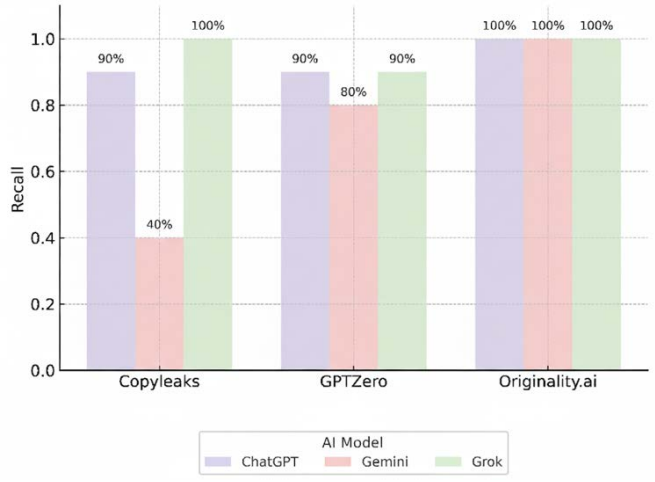


Figure 3. Comparative chart of tool recall according to the LLM used for text generation.

As for Grok and ChatGPT, both models prove to be more easily detectable. GPTZero achieves a recall of 90% with both, although the analysis of confidence levels reveals important nuances: the tool shows high confidence in the TPs produced with Grok, whereas in 40% of ChatGPT texts, the confidence is medium or low, weakening the result. Likewise, it misclassifies one of the ChatGPT texts with high confidence, in contrast to the FN it commits with Grok, which is reported with low confidence. This detail aggravates the interpretation of error from a forensic perspective, where not only the accuracy rate matters but also the subjective reliability of the system.

By contrast, Originality.ai demonstrates outstanding consistency across the three models: it correctly detects all texts without exception or loss of confidence, reinforcing its robustness against inter-model variability.

In summary, the performance of the tools differs notably both quantitatively and qualitatively. Originality.ai offers the best coverage and the greatest stability, being especially reliable for detecting artificial texts generated by diverse LLMs. GPTZero emerges as a more conservative option, particularly advisable when minimizing false positives in human texts is a priority. Copyleaks, on the other hand, shows more erratic and error-prone behavior, both in terms of recall and precision, which compromises its practical applicability in contexts where reliability is a critical requirement.

These results confirm that the choice of detector should neither be neutral nor automatic; one must take into account not only global metrics but also the type of text analyzed, the AI model involved, and the purpose of the analysis. In particular, errors made with high confidence or those systematically associated with a specific LLM (such as Gemini) constitute a risk factor in academic, legal, or publishing environments where the consequences of misclassification can be especially serious.

In addition to the analysis of hits and errors, the level of confidence with which the tools issue their predictions has also been considered, as this parameter is especially relevant in forensic, academic, or publishing contexts. An error made with high confidence not only increases the risk of false attributions but also undermines the credibility of the system by conveying a false sense of certainty.

## 4.2 Hybrid texts

One of the main limitations observed in current AI text detection tools is their limited ability to recognize cases of shared authorship or cross-intervention between humans and machines. In this study, we included, on the one hand, texts generated by AI and subsequently modified by humans, and on the other, human texts that were edited, rewritten, or reformulated by AI.
Both scenarios represent real situations in academic, professional, or forensic settings, where partial collaboration or substitution between human agents and artificial systems is increasingly frequent. Separating these two variants not only allows comparison of the overall recall of detectors on mixed texts but also enables analysis of whether the order and direction of intervention affect their performance.

### 4.2.1 AI + H: human intervention on AI-generated text

For this category, the AI-generated texts were modified through three distinct types of human intervention: insertion of sentences, changes in punctuation, and introduction of grammatical errors. The aim of this section is to assess whether detection tools are able to identify the underlying artificial trace even after human intervention, and to what extent their performance is altered by the type of modification or by the original generative model. The results indicate that the specific type of human intervention does not significantly alter detector performance. As shown in Figure 4, the three types of alteration produce very similar results for each tool. The results, however, appear to be more closely related to the LLM originally used in the production of the texts.

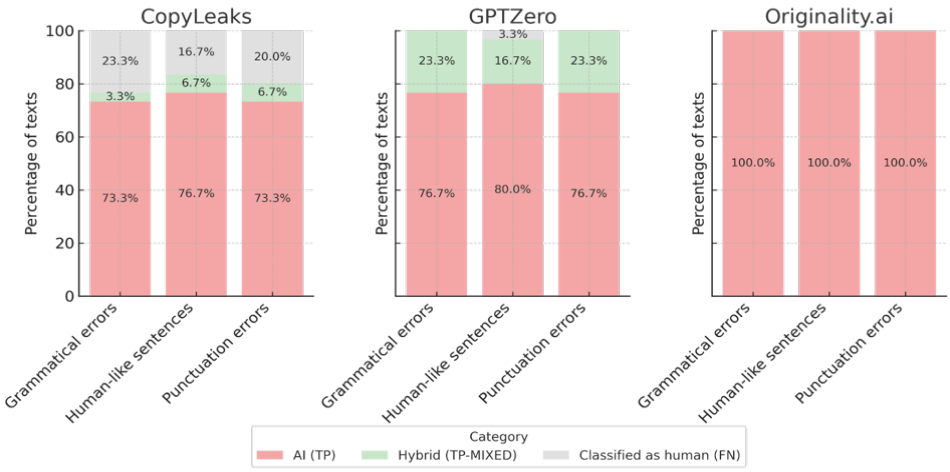


Figure 4. Comparative chart of AI texts modified by humans, by tool and type of modification.

In terms of synthetic detection performance, Originality.ai stands out for its effectiveness: it detects AI involvement in 100% of the texts in this category, even after human modifications. In all cases, the tool maintains high confidence levels consistent with those obtained before the alterations, indicating strong model stability against minor or stylistic manipulations. Only in one case was a slight decrease in confidence observed (from 100% to 98%), without affecting the final classification.

GPTZero, in turn, also shows high performance, with a detection rate of 98.9% in this category. It makes only one false negative, suggesting that its recall remains high, although somewhat less stable than that of Originality.ai. By contrast, Copyleaks exhibits clearly lower performance: it makes a total of 18 FNs in this category, meaning that it fails to detect AI involvement in 20% of the modified texts. This behavior reflects the system's limited ability to recognize artificial traces once they have been partially masked by human intervention.

A more detailed analysis, however, reveals that in the case of Copyleaks, the drop in performance is not due exclusively to human modifications, since only 4 texts represent new errors compared to the classification of purely AI texts. In fact, 77.8% of the errors committed by this tool in this category had already occurred in the previous stage (unmodified AI texts). Therefore, it can be concluded that its poor performance mainly reflects a structural limitation of the detector with certain generative models, rather than a direct effect of human modifications.

In this sense, the distribution of errors according to the generative model confirms a trend already observed in the binary analysis. Texts originally generated by Gemini are significantly more difficult for Copyleaks to detect, accounting for 14 of its 18 errors in this subcategory. The remaining 4 errors correspond to texts generated by ChatGPT, while none occur with Grok.

Taken together, these results indicate that superficial human intervention on artificial texts does not critically affect the performance of more advanced detectors such as Originality.ai and GPTZero. However, it does highlight certain limitations in systems such as Copyleaks, whose precision is unevenly affected by the generative model (especially Gemini) and by the presence of human modifications. Although in our study its error rate decreases slightly when moving from purely artificial texts to modified texts (from 23.3% to 20%), this decrease does not reflect a consistent improvement but rather an unstable response, contrasting with the greater robustness observed in other detectors.

Moreover, the stability of results after modifications to AI texts may be due to relatively limited human intervention. Although not precisely quantified, this could be a factor contributing to the fact that detectors largely maintain their initial classifications after the changes.

In addition to detecting AI involvement after human modifications, the tools were also evaluated for their ability to identify the hybrid nature of these texts. This task is particularly demanding, as it requires not only detecting the presence of artificial elements but also the trace of subsequent human intervention. As shown in Figure 5, the results in this dimension were notably lower than those obtained in the binary AI/Human classification.

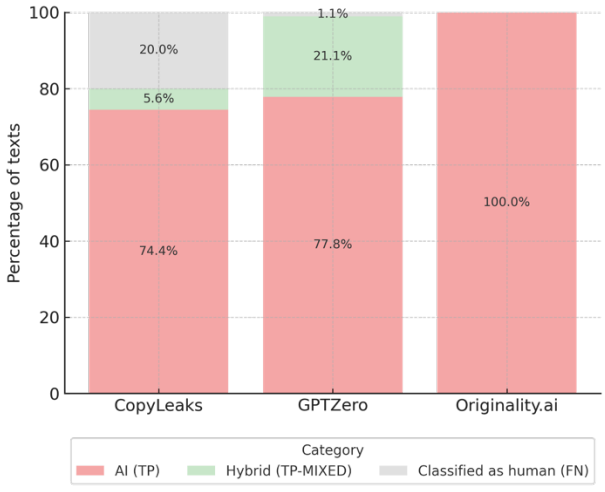


Figure 5. Comparative chart of detectors according to their classification of AI texts with human modifications.

GPTZero was the only tool that produced a significant number of classifications as *Mixed*, with a total of 19 texts (21.1%). However, a qualitative analysis of these labels reveals important limitations. In 18 out of the 19 cases, GPTZero described the text as *originally human and edited by AI*, a formulation that completely inverts the actual direction of the process, since all texts in this category were initially generated by AI and subsequently modified by humans. This inversion, although formally labeled as *Mixed*, implies a misinterpretation of the type of intervention, which undermines the reliability of the classification from a forensic or academic perspective.

### 4.2.2 H + AI: artificial intervention on human text

This subsection focuses on the second hybrid text modality analyzed: those initially written by humans and subsequently edited or reformulated by AI tools. This type of intervention represents an increasingly common scenario, both in academic and professional contexts, where users turn to generative systems to rewrite, correct, or improve their own text.
For the H texts modified by the three AI models, the same prompt was used (see Section 3.2). The response of each model to this prompt was different: Grok and ChatGPT rewrote the text almost entirely in most cases, whereas Gemini's modifications were more respectful of the original content. The aim was to observe the extent to which detection tools are able to identify the presence of artificial intervention in an otherwise genuinely human text.

**Partial rewriting with Gemini**

As shown in Figure 7, the human texts modified by Gemini generated the largest number of errors in this category: 10 of the 11 FNs (90.9%) made by the detectors on H + AI texts are concentrated in this submodality. Within this set, Originality.ai was once again the most accurate tool, with 90% accuracy and a single false negative. Copyleaks ranked second, with an accuracy rate of 80%. By contrast, GPTZero showed the lowest performance, with only 30% accuracy, while in 70% of cases it wrongly attributed authorship exclusively to humans.

**Total rewriting with Grok and ChatGPT**

In the texts extensively modified by Grok and ChatGPT, the most reliable detectors were Originality.ai and GPTZero. Originality.ai identified AI intervention in 100% of cases, while GPTZero classified 75% as AI production and the remaining 25% as hybrid. Although GPTZero's explicit accuracy rate for *Mixed* is low, it represents the best performance among the three detectors in this subcategory. Copyleaks, for its part, was the only detector to commit a false negative, classifying as human a text completely rewritten by Grok. This error is particularly serious due to the clarity of the case and reinforces the system's low reliability when faced with extensive artificial interventions.
With regard to the identification of hybrid nature, results with this label were scarce. Copyleaks classified only one text as *Mixed*, and this result is questionable, since the same text had been incorrectly classified as AI in its purely human version. After ChatGPT's total intervention, it was then labeled as *Mixed*, which suggests an internal oscillation rather than a conscious detection of hybridity.

This finding is particularly significant: GPTZero's 7 FNs in this subcategory account for 58.3% of all such errors, even though H + AI texts represent only 16.7% of the corpus. This imbalance confirms that GPTZero has particular difficulty detecting Gemini's intervention, both when it rewrites a human text and with purely synthetic texts (see also Section 4.1). Added to this is the fact that in most of these cases, the errors are issued with low confidence levels, indicating a lack of internal certainty that may mitigate the seriousness of the failure from a forensic perspective. The tool itself acknowledges that lower confidence implies a higher probability of error .

Finally, it should be noted that virtually all of the hybrid classification errors (TP-Mixed) issued by GPTZero in this category correspond to texts generated by Gemini. This suggests that the architecture of the generative model and the depth of intervention significantly affect detector performance, even in cases where the tool is correct in binary terms, as observed in this study.

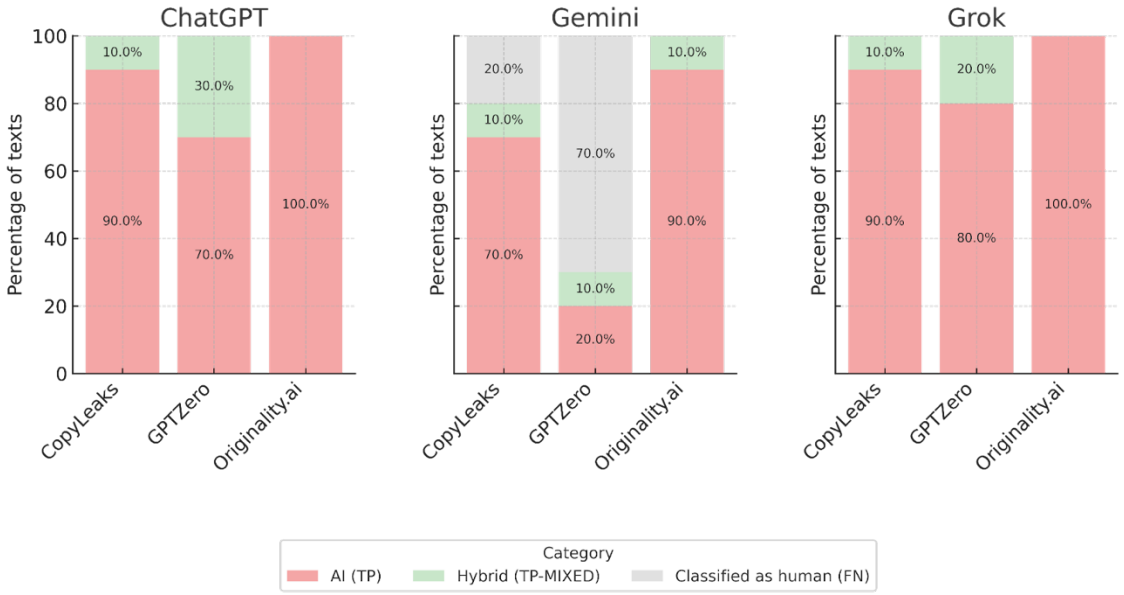


Figure 7. Comparative Chart of Human Texts Enhanced by AI, by Tool and LLM.

# 5 Discussion and conclusions

The present study achieved its main objective: to comparatively evaluate the reliability of three commercial AI detection tools (Originality.ai, GPTZero, and Copyleaks) applied to human, artificial, and mixed texts in Spanish. The results confirm that detector performance varies significantly depending on the authorship modality and the generative model (LLM) involved, as well as on the degree of hybridity of the text. In general terms, Originality.ai was the most reliable tool, with 100% accuracy in texts fully generated by AI and 90% in fully human texts, yielding an overall recall of 99.3%. However, the false positive rate for human texts was 10%. The metrics reflect excellent performance in AI detection, but a moderate FPR, which is problematic in forensic settings. This finding is consistent with previous studies such as [1, 2, 3], which had already highlighted the recall of Originality.ai compared to other detectors, particularly in binary classification tasks.

Secondly, GPTZero showed acceptable performance with human texts but suffered a significant decline with texts generated by Gemini and also with hybrid texts involving partial intervention. This pattern reproduces the limitations already noted in works such as [4], which emphasize the vulnerability of many detectors when confronted with models underrepresented in their training data. Although GPTZero's internal benchmark for Spanish [53] reports 98.9% accuracy, 98.2% recall, 0.4% false positives, and an F1-score of 99.1%, our data reveal less uniform behavior when multi-authored or cross-intervention texts are evaluated.

Finally, Copyleaks was the tool with the worst overall performance. Despite its commercial claim of exceeding 99% precision, in this study it committed numerous errors, especially in misclassifying human texts as AI-generated or in failing to detect artificial intervention in hybrid texts. Its errors are concentrated in texts generated by Gemini, reinforcing the hypothesis of a training bias toward predominant models such as GPT-4. This divergence between advertised and observed performance calls into question the practical reliability of Copyleaks, in line with the results of [40, 43], who also warn of variability in the effectiveness of these tools depending on the context of use.

Taken together, the findings not only reinforce evidence regarding biases and limitations of current detectors but also introduce new variables of analysis—such as the direction of textual hybridity and sensitivity by generative model—that help to better understand their possibilities and limits. This work also highlights the risks of automating attribution processes without expert mediation and proposes analytical tools from forensic linguistics to contextualize the results. Even the most accurate systems commit serious errors, such as high-confidence false positives or incorrect attributions in hybrid texts, which, combined with the lack of explainability, compromise their use in legal or academic contexts.

Looking ahead, three main limitations should be underlined: the dependence on specific versions of detectors, the opacity of the internal functioning of models such as GPTZero, and the absence of systematic treatment of hybrid texts. These limitations open up lines of research that include the evaluation of tools in multilingual or translated contexts, the incorporation of emerging detectors, the comparison between expert linguistic judgments and automatic systems, the expansion of the corpus to other genres and registers, and the qualitative analysis of detector-generated reports. In short, this work contributes to a critical, explanatory, and ethically informed approach to authorship analysis in the age of artificial intelligence.

In conclusion, this study demonstrates that while commercial AI detectors such as Originality.ai, GPTZero, and Copyleaks can achieve notable performance in binary classification, their effectiveness decreases substantially in hybrid authorship scenarios. The results highlight the importance of considering not only global metrics but also the impact of specific generative models and the direction of human–AI interaction. These findings reinforce the need for cautious, expert-mediated interpretation of detector outputs, especially in forensic and academic contexts where the risks of misclassification are critical. Future research should prioritize multilingual evaluation, transparency of detection models, and systematic treatment of hybrid texts, in order to develop more reliable and explainable tools for authorship analysis in the era of artificial intelligence.

# References


[1] Arslan Akram. 2023. An empirical study of AI generated text detection tools. *arXiv preprint arXiv:2310.01423*. DOI: https://doi.org/10.48550/arXiv.2310.01423

[2] William H. Walters. 2023. The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science, 7*(1). DOI: https://doi-org.sire.ub.edu/10.1515/opis-2022-0158

[3] Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito & Chris Callison-Burch. 2024. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. *arXiv preprint arXiv:2405.07940*. DOI: https://doi.org/10.48550/arXiv.2405.07940

[4] Valentina Bellini, Federico Semeraro, Jonathan Montomoli, Marco Cascella & Elena Bignami. 2024. Between human and AI: assessing the reliability of AI text detection tools. *Current Medical Research and Opinion, 40*(3), 353-358. DOI: https://doi.org/10.1080/03007995.2024.2310086

[5] Rong Zheng, Yi Qin, Zan Huang & Hsinchun Chen. 2003. Authorship analysis in cybercrime investigation. *International conference on intelligence and security informatics*. Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: https://doi-org.sire.ub.edu/10.1007/3-540-44853-5_5

[6] Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, *22*(3), 251-270. DOI: https://doi-org.sire.ub.edu/10.1093/llc/fqm020

[7] María Teresa Turell & Núria Gavaldà. 2013. Towards an Index of Idiolectal Similitude (Or Distance) In Forensic Authorship Analysis. *Journal of Law and Policy*, *21*(2), 10.

[8] Janet Ainsworth & Patrick Juola. 2018. Who wrote this: Modern forensic authorship analysis as a model for valid forensic science. *Wash. UL Rev.*, *96*, 1159.

[9] Elena Garayzábal Heinze, Sheila Queralt Estévez & Mercedes Reigosa Riveiros. 2019. *Fundamentos de la linguistica forense*. Síntesis.

[10] Tim Grant. 2022. *The Idea of Progress in Forensic Authorship Analysis*. Cambridge Elements.

[11] Serena Fariello, Giuseppe Fenza, Flavia Forte, Mariacristina Gallo & Martina Marotta. 2025. Distinguishing Human From Machine: A Review of Advances and Challenges in AI-Generated Text Detection. *International Journal of Interactive Multimedia & Artificial Intelligence* 9.3. DOI: https://doi.org/10.9781/ijimai.2024.12.002

[12] Boston University AI Task Force. 2024. *Report on Generative AI in Education and Research*. Boston University.

[13] María Camila Bernal Carvajal. 2023. *ChatGPT: Modalidades de Fraude, Métodos de Detección y Estrategias Antiplagio a partir de Testimonios Docentes*. Lenguas Modernas-Virtual. EAN University.

[14] Feng Kevin Jiang & Ken Hyland. 2025. Rhetorical distinctions: Comparing metadiscourse in essays by ChatGPT and students. *English for Specific Purposes* 79. 17-29. DOI: https://doi.org/10.1016/j.esp.2025.03.001

[15] Feng Kevin Jiang & Ken Hyland. 2025. Metadiscursive nouns in academic argument: ChatGPT vs student practices. *Journal of English for Academic Purposes* 75. DOI: https://doi.org/10.1016/j.jeap.2025.101514

[16] J. Elliott Casal & Matt Kessler. 2023. Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, *2*(3). DOI: https://doi.org/10.1016/j.rmal.2023.100068

[17] Muneera Bano, Didar Zowghi, Jon Whittle, Liming Zhu & Andrew Reeson. 2025. A Qualitative Study of User Perception of M365 AI Copilot. *arXiv preprint arXiv:2503*. DOI: https://doi.org/10.48550/arXiv.2503.17661

[18] Sandra Mitrovic, Davide Andreoletti & Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text." *arXiv preprint arXiv:2301*. DOI: https://doi.org/10.48550/arXiv.2301.13852

[19] Soohyeon Choi, Yong Kiam Tan, Mark Huasong Meng, Mohamed Ragab, Soumik Mondal, David Mohaisen & Khin Mi Mi Aung. 2025. I can find you in seconds! leveraging large language models for code authorship attribution. *arXiv preprint arXiv:2501*. DOI: https://doi.org/10.48550/arXiv.2501.08165

[20] Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi & Paolo Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains." *arXiv preprint arXiv:2309*. DOI: https://doi.org/10.48550/arXiv.2309.11285

[21] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Chenxi Whitehouse, Alham Fikri Aji, Nizar Habash, Iryna Gurevych & Preslav Nakov. 2024. Semeval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. *arXiv preprint arXiv:2404.14183*. DOI: https://doi.org/10.48550/arXiv.2404.14183

[22] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozi.re, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin Edouard Grave & Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. DOI: https://doi.org/10.48550/arXiv.2302.13971

[23] Jonathan Gillham. 2024. Can Grok AI content be detected? Originality.AI. https://originality.ai/blog/can-grok-ai-content-be-detected

[24] Jonathan Gillham. 2024. Can Mixtral AI content be detected? Originality.AI. https://originality.ai/blog/can-mixtral-ai-content-be-detected

[25] Srinivasa Rao Bogireddy & Nagaraju Dasari. 2024. Comparative analysis of ChatGPT-4 and LLaMA: Performance evaluation on text summarization, data analysis, and question answering. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* 1-7, IEEE. DOI: https://doi.org/10.1109/ICCCNT61001.2024.10725662

[26] Vamsi Krishna Uppalapati & Deb Sanjay Nag. 2024. A comparative analysis of AI models in complex medical decision-making scenarios: evaluating ChatGPT, Claude AI, Bard, and Perplexity. *Cureus*, *16*(1). DOI: 10.7759/cureus.52485

[27] Jess Sawyer. 2025. Grok AI: 2025 statistics and facts about Elon Musk's AI challenger to ChatGPT. Originality.AI. https://originality.ai/blog/grok-ai-statistics

[28] Pusheng Xu, Yue Wu, Kai Jin, Xiaolan Chen, Mingguang He & Danli Shi. 2025. Deepseek-r1 outperforms gemini 2.0 pro, openai o1, and o3-mini in bilingual complex ophthalmology reasoning. *Advances in Ophthalmology Practice and Research*. DOI: https://doi.org/10.1016/j.aopr.2025.05.001

[29] Alex Adam, Edwin Thomas & Vivienne Chen. 2025. *GPTZero 2025 Benchmarks: How we detect ChatGPT o1*. GPTZero. https://gptzero.me/news/gptzero-o1-benchmarking/

[30] Hooman H. Rashidi, Brandon D. Fennell, Samer Albahra, Bo Hub & Tom Gorbett. 2023. The ChatGPT conundrum: Human-generated scientific manuscripts misidentified as AI creations by AI text detection tool. *Journal of Pathology Informatics*, *14*. DOI: https://doi.org/10.1016/j.jpi.2023.100342

[31] Rafael Rivera Soto, Kailin Koch, Aleem Khan, Barry Chen, Marcus Bishop & Nicholas Andrews. 2024. Few-shot detection of machine-generated text using style representations. *arXiv preprint arXiv:2401.06712*. DOI: https://doi.org/10.48550/arXiv.2401.06712

[32] Vittoria Dentella, Weihang Huang, Silvia A. Mansi, Jack Grieve & Evelina Leivada. 2025. ChatGPT-generated texts show authorship traits that identify them as non-human. *arXiv preprint arXiv:2508.16385*. https://doi.org/10.48550/arXiv.2508.16385

[33] Mitchell Clark. 2023. *ChatGPT's creator made a free tool for detecting AI-generated text*. The Verge. https://www.theverge.com/2023/1/31/23579942/chatgpt-ai-text-detection-openai-classifier

[34] Emilia David. 2023. *OpenAI can't tell if something was written by AI after all*. The Verge. https://www.theverge.com/2023/7/25/23807487/openai-ai-generated-low-accuracy

[35] Sebastian Gehrmann, Hendrik Strobelt & Alexander M. Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*. DOI: https://doi.org/10.48550/arXiv.1906.04043
[36] Seyhan Canyakan. 2025. Comparative accuracy of AI-based plagiarism detection tools: an enhanced systematic review. *Journal of AI, Humanities and New Ethics*, 5-18. DOI: https://doi.org/10.5281/zenodo
[37] Vitalii Fishchuk & Daniel Braun. 2024. Robustness of generative AI detection: adversarial attacks on black-box neural text detectors. *International Journal of Speech Technology*, *27*(4), 861-874. DOI: https://doi-org.sire.ub.edu/10.1007/s10772-024-10144-2
[38] Frederick M. Howard, Anran Li, Mark F. Riffon, Elizabeth Garrett-Mayer & Alexander T. Pearson. 2024. Characterizing the increase in artificial intelligence content detection in oncology scientific abstracts from 2021 to 2023. *JCO Clinical Cancer Informatics*, *8*. DOI: https://doi.org/10.1200/CCI.24.00077
[39] Sujita Kumar Kar, Teena Bansal, Sumit Modi & Amit Singh. 2025. How sensitive are the free AI-detector tools in detecting AI-generated texts? A comparison of popular AI-detector tools. *Indian Journal of Psychological Medicine*, *47*(3), 275-278. DOI: https://doi-org.sire.ub.edu/10.1177/02537176241247934
[40] Pablo Picazo-Sanchez & Lara Ortiz-Martin. 2024. Analysing the impact of ChatGPT in research. *Applied Intelligence*, *54*(5), 4172-4188. DOI: https://doi.org/10.1007/s10489-024-05298-0
[41] Kathleen C. Fraser, Hillary Dawkins & Svetlana Kiritchenko. 2025. Detecting ai-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research*, *82*, 2233-2278. DOI: https://doi.org/10.1613/jair.1.16665
[42] Jonathan Gillham. 2025. AI Detection Accuracy Studies — Meta-analysis of 10 studies. Originality.ai. https://originality.ai/blog/ai-detection-studies-round-up
[43] Ayat A. Najjar, Huthaifa I. Ashqar, Omar A. Darwish & Eman Hammad. 2025. Detecting AI-Generated Text in Educational Content: Leveraging Machine Learning and Explainable AI for Academic Integrity. *arXiv preprint arXiv:2501.03203*. DOI: https://doi.org/10.48550/arXiv.2501.03203
[44] Shushanta Pudasaini, Luis Miralles-Pechuán, Marisa Llorens Salvador & David Lillis. 2025. Benchmarking AI Text Detection: Assessing Detectors Against New Datasets, Evasion Tactics, and Enhanced LLMs. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)* 68-77.
[45] Jingyi Liu, Youyan Nie & Bee Leng Chua. 2024. Generative AI in assessment: AI detectors and implications for practice. DOI: https://doi.org/10.21203/rs.3.rs-4540908/v1
[46] Zijie Zeng, Shiqi Liu, Lele Sha, Zhuang Li, Kaixun Yang, Sannyuya Liu, Dragan Gasevic & Guanliang Chen. 2024. Detecting ai-generated sentences in human-ai collaborative hybrid texts: Challenges, strategies, and insights. *arXiv preprint arXiv:2403.03506*. DOI: https://doi.org/10.48550/arXiv.2403.03506
[47] Lara Alonso Simón, Ana María Fernández-Pampillón Cesteros, Marianela Fernández Trinidad & Manuel Márquez Cruz. 2024. ¿Tienen GPT-3.5 y GPT-4 un estilo de escritura diferente del estilo humano? Un estudio exploratorio para el español. *Revista Electrónica de Lingüística Aplicada*, *23*(1). DOI: https://doi.org/10.58859/rael.v23i1.666
[48] Sergio E. Zanotto & Segun Aroyehun. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models. *arXiv preprint arXiv:2412.03025*. DOI: https://doi.org/10.48550/arXiv.2412.03025
[49] Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut & Lorna Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), 1-39. DOI: https://doi.org/10.1007/s40979-023-00146-z
[50] Sheila Queralt. 2015. Estudio piloto para la evaluación de evidencias lingüísticas en la comparación forense de textos mediante distribuciones poblacionales y relaciones de verosimilitudes. PHD thesis, Pompeu Fabra University (UPF).
[51] Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology* 60.3, 538-556. DOI: https://doi.org/10.1002/asi.21001
[52] Graeme Hirst & Olga Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing* 22.4. 405-417. DOI: https://doi-org.sire.ub.edu/10.1093/llc/fqm023
[53] Emily Napier & Alex Adam. 2025. Behind the Scenes: Multilingual Detection. GPTZero. https://gptzero.me/news/behind-the-scenes-multilingual-detection/
[54] Sheila Queralt. 2024. Los retos de la lingüística forense en la era de la IA. *Lengua y Sociedad*, *23*(2), 1099-1118. DOI: http://dx.doi.org/10.15381/lengsoc.v23i2.29462